BIOMETRICS 58, 502–509
September 2002

Analysis of Survival Data from Case–Control Family Studies

Joanna H. Shih

Division of Cancer Treatment and Diagnosis, National Cancer Institute,
6130 Executive Boulevard, Bethesda, Maryland 20892-7434, U.S.A.
email: jshih@mail.nih.gov

and

Nilanjan Chatterjee

Division of Cancer Epidemiology and Genetics, National Cancer Institute,
6120 Executive Boulevard, Bethesda, Maryland 20892, U.S.A.
email: chattern@mail.nih.gov

SUMMARY. In case–control family studies with survival endpoint, age of onset of diseases can be used to assess the familial aggregation of the disease and the relationship between the disease and genetic or environmental risk factors. Because of the retrospective nature of the case–control study, methods for analyzing prospectively collected correlated failure time data do not apply directly. In this article, we propose a semiparametric quasi-partial-likelihood approach to simultaneously estimate the effect of covariates on the age of onset and the association of ages of onset among family members that does not require specification of the baseline marginal distribution. We conducted a simulation study to evaluate the performance of the proposed approach and compare it with the existing semiparametric ones. Simulation results demonstrate that the proposed approach has better performance in terms of consistency and efficiency. We illustrate the methodology using a subset of data from the Washington Ashkenazi Study.

KEY WORDS: Age of onset; Case–control family studies; Copula models; Correlated failure times; Semiparametric models.

1. Introduction

Case–control family studies have been used to assess the familial aggregation of a disease and the relationship between the disease and genetic or environmental risk factors. Such a case–control family study identifies a sample of cases who develop the disease of interest and an independent sample of controls who are free of the disease at the time of ascertainment. From each identified individual, hereafter called proband, information collected includes disease outcomes (age of onset or age of censoring) and risk factors of the proband and the relatives. For rare diseases, the case–control design provides an efficient way to ascertain a large number of cases in a short period of time.

It is standard in the survival analysis to treat the incidence and age of onset as one composite disease outcome. Then the familial aggregation captures not only the correlation of the incidence but also the correlation between the ages of onset. Chatterjee and Shih (2001) develop a bivariate model separating the incidence and age of onset of disease. In this article, we do not separate the two types of expression for a disease but take the standard approach of using the composite disease outcome. Several statistical issues arise when analyzing data collected from such a case–control family study. Methods that fail to account for the available age of onset may lose an important feature of the data. Strong correlation in early ages among family members, e.g., often indicates genetic predisposition. In addition, individuals free of disease at the time of ascertainment may develop the disease at a later time. If the information on age of onset and the effect of censoring are ignored and the outcomes are simply classified as diseased or nondiseased, then the degree of familial aggregation may be underestimated (MacLean et al., 1990) and the estimators of the covariate effect on the onset of the disease may lose efficiency. Further, ages of onset of the relatives are correlated due to the genes and environments they share. We need to account for such a correlation when studying the association between the risk factors and disease and use it to assess the familial aggregation of the disease. Incorporating the age-of-onset information, however, may be challenging due to the retrospective sampling of the probands. Recent advancement of methodology on the analysis of correlated survival time data (Shih and Louis, 1995; Prentice and Hsu, 1997), on the other hand, can be used to estimate familial association but requires that the data be collected from prospective studies and form, at least approximately, a random sample of the population and thus cannot be applied directly to the case–

control family studies. Prentice and Breslow (1978) propose a method for analyzing matched survival time data from case–control studies. Though the method can be used to estimate the effect of risk factors on survival, it does not estimate the familial association.

Recently, Li, Yang, and Schwartz (1998) proposed a parametric likelihood approach that addresses the above-mentioned issues when analyzing data collected from case–control family studies. The purpose of this article is to propose a semiparametric counterpart for studying the association between the disease onset and risk factors and assessing the familial association in age of onset of disease. With a specified model for the bivariate failure time distribution, we propose an iterative procedure to estimate the marginal baseline distribution nonparametrically. For a given marginal baseline distribution, we propose a quasi-partial-likelihood approach to estimate the parameters that measure the relationship between risk factors and disease outcome and strength of familial association.

The rest of the article is organized as follows. In Section 2, we review the likelihoods considered by Whittemore (1995) and Li et al. (1998). In Section 3, we propose an iterative method for estimating the marginal baseline distribution and a quasi-partial-likelihood for estimating the covariate effects and familial association. In Sections 4 and 5, we use simulation and an analysis of data from the Washington Ashkenazi Study (WAS) to illustrate the proposed method. A discussion follows in Section 6 to conclude the article.

2. A Likelihood Approach for Age-of-Onset Data from Case–Control Family Studies

Consider a matched case control family study where one case proband is matched in age with one control proband. Each matched set contains one case family and one control family, and there are a total of n matched set. Let $(\boldsymbol{T}_i, \boldsymbol{\Delta}_i) = \{(t_{i0}, \ldots, t_{im_i}), (\delta_{i0}, \ldots, \delta_{im_i})\}$, $i = 1, \ldots, 2n$ denote the disease outcomes and $\boldsymbol{Z}_i = (\boldsymbol{z}_{i0}, \ldots, \boldsymbol{z}_{im_i})$ the associated covariates for the ith family of size $m_i + 1$, with the first component in the vectors corresponding to the proband. The binary variable δ_{ij} indicates whether the individual developed the disease ($\delta_{ij} = 1$) or not ($\delta_{ij} = 0$), and t_{ij} denotes the age of onset if $\delta_{ij} = 1$ and censoring time if $\delta_{ij} = 0$. Let the first n families be case families and the remaining families be control families. By design, $\delta_{i0} = 1$, $\delta_{i+n,0} = 0$, and $t_{i0} = t_{i+n,0}$, $i = 1, \ldots, n$. The retrospective likelihood for the case–control family study is

$$L = \prod_{i=1}^{2n} \Pr\left\{\left(\boldsymbol{T}_i^{-1}, \boldsymbol{\Delta}_i^{-1}\right), \boldsymbol{Z}_i \mid (t_{i0}, \delta_{i0})\right\},$$

where the superscript -1 denotes a vector with its first component removed. The likelihood can be factored as

$$L = \prod_{i=1}^{2n} \Pr\{\boldsymbol{z}_{i0} \mid (t_{i0}, \delta_{i0})\} \Pr\left\{\boldsymbol{Z}_i^{-1} \mid \boldsymbol{z}_{i0}, (t_{i0}, \delta_{i0})\right\} \times \Pr\left\{\left(\boldsymbol{T}_i^{-1}, \boldsymbol{\Delta}_i^{-1}\right) \mid \boldsymbol{Z}_i, (t_{i0}, \delta_{i0})\right\}. \quad (1)$$

Now we make the reproducibility assumption for marginal models (Whittemore, 1995), i.e., $\Pr\{(t_{ij}, \delta_{ij}) \mid \boldsymbol{Z}_i\} = \Pr\{(t_{ij}, \delta_{ij}) \mid \boldsymbol{z}_{ij}\}$. Because, under this assumption, $\boldsymbol{Z}_i^{-1}$ is conditionally independent of (t_{i0}, δ_{i0}) given $\boldsymbol{z}_{i0}$, the second factor in the likelihood expression (1) simplifies to $\Pr\{(\boldsymbol{Z}_i)^{-1} \mid \boldsymbol{z}_{i0}\}$ and hence can be ignored because it does not depend on the parameters of interest. Heretofore, in the likelihood expression (1), we will denote the first factor by L_1 and the third factor by L_2. The decomposition of L into the product of L_1 and L_2 implies that an individual's covariates do not affect another family members' ages of onset, but they may affect the familial association of ages of onset. For example, if two family members have identical covariates, then their ages of onset are likely to be more similar than if they have very different covariates.

Although the above retrospective likelihood conditions on the proband's data, it does not take into account the matching of age of onset. Li et al. (1998) account for the matching by replacing L_1 with the conditional likelihood of Prentice and Breslow (1978). Also, they use the Clayton model (Clayton, 1978) to specify the multivariate distribution of ages of onset for L_2. Specifically, assume the marginal distribution of age of onset for each individual follows a proportional hazards model, with the hazard function given by

$$\lambda(t \mid \boldsymbol{z}) = \lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}). \quad (2)$$

Suppose there are k distinct ages of onset among the case probands and, at the ith age, there are k_i case probands and l_i control probands (for our case of one-to-one matching, $k_i = l_i$) selected with covariates $\boldsymbol{z}_{10}, \ldots, \boldsymbol{z}_{k_i,0}$ and $\boldsymbol{z}_{k_i+1,0}, \ldots, \boldsymbol{z}_{k_i+l_i,0}$, respectively. Then, at the ith age, the probability that covariates $\boldsymbol{z}_{10}, \ldots, \boldsymbol{z}_{k_i0}$ correspond to the cases given the $k_i + l_i$ covariates under the proportional hazards model is

$$\frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\sum_{j \in R(k_i, l_i)} \exp(\boldsymbol{\beta}' \boldsymbol{s}_j)}, \quad (3)$$

where $\boldsymbol{s}_i = \boldsymbol{z}_{10} + \cdots + \boldsymbol{z}_{k_i,0}$, $\boldsymbol{s}_j = \boldsymbol{z}_{j_1} + \cdots + \boldsymbol{z}_{j_{k_i}}$, and $R(k_i, l_i)$ is the set of all subsets of size k_i from set $\{1, \ldots, k_i + l_i\}$. The conditional likelihood of Prentice and Breslow (1978) is the product of (3) over all k distinct ages, given by

$$L_1^c(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\sum_{j \in R(k_i, l_i)} \exp(\boldsymbol{\beta}' \boldsymbol{s}_j)}.$$

The other component in (1), L_2, requires specification of the joint distribution of the age of onset for the relatives and proband. The joint survival function from the Clayton model is given by

$$\Pr(T_0 > t_0, T_1 > t_1, \ldots, T_l > t_l) = \left[\sum_{j=0}^{l} S_j(t_j)^{1-\theta} - (l-1)\right]^{1/(1-\theta)},$$

where $\theta = \exp(\alpha) \geq 1$ is the association parameter that has the cross-ratio interpretation (Oakes, 1989) and S_j is the marginal survival pertaining to the jth failure time. Unity of θ corresponds to independence, and a value greater than one indicates positive association. Under the proportional hazards model (2), $S_j(t) = \exp\{-\Lambda_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{z}_j)\}$, where $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$. Assume the censoring time of each individual is

independent of the age of onset. Under the Clayton model, L_2 is given by

$$
\begin{aligned}
&L_2(\beta,\gamma,\alpha)\\
&\quad=\prod_{i=1}^{2n}\left[\prod_{j=1}^{d_i}\{\theta(j-1)+2-j\}\right]\\
&\qquad\times\left[\sum_{j=0}^{m_i}\exp\{-\Lambda_0(t_{ij};\gamma)(1-\theta)\exp(\beta' z_{ij})\}\right.\\
&\qquad\qquad\left.-m_i\right]^{1/(1-\theta)-d_i}\\
&\qquad\times\left[\prod_{j=0}^{m_i}[\exp\{-\Lambda_0(t_{ij};\gamma)(1-\theta)\exp(\beta' z_{ij})\}\right.\\
&\qquad\qquad\left.\times\lambda_0(t_{ij};\gamma)\exp(\beta' z_{ij})]^{\delta_{ij}}\right]\\
&\qquad\times\exp\{\Lambda_0(t_{i0};\gamma)\exp(\beta' z_{i0})\}\\
&\qquad\times\{\exp(\beta' z_{i0})\lambda_0(t_{i0};\gamma)\}^{-\delta_{i0}},
\end{aligned}
\tag{4}
$$

where γ are the parameters associated with the marginal baseline distribution and $d_i=\Sigma_{j=0}^{m_i}\delta_{ij}$. We obtain L_2 by deriving the joint likelihood for the relatives and proband, assuming the Clayton model, and then conditioning on $T_{i0}=t_{i0}$ for the case proband (or $T_{i0}>t_{i0}$ for the control proband). See the Appendix.

Strictly speaking, the likelihood considered by Li et al. (1998) is not a full but pseudo-likelihood. Nevertheless, L_1^c accounts for matching and is a valid likelihood for including the contribution of the data from probands to the estimation of β. The estimates of (β,γ,α) are obtained by maximizing $L_1^cL_2$ with respect to (β,γ,α).

3. Proposed Method

The likelihood approach in the previous section requires specification of the marginal baseline distribution up to a finite number of parameters (γ). In this section, we present a new method for estimating β, α, and $\Lambda_0(\cdot)$ nonparametrically.

3.1 *Model*

We begin by assuming covariates are not present. Consider the hazard function for the age of onset of a relative given the disease outcome of the proband, denoted by $\lambda(t\mid t_0,\delta_0)$. Assume the ages of onset for the proband and relative come from an absolutely continuous distribution. Then $\lambda(t\mid t_0,\delta_0)$, by definition, can be represented by

$$\lambda(t\mid t_0,\delta_0)=\lambda(t\mid t_0,0)\psi(t,t_0)^{\delta_0},$$

where $\psi(t,t_0)$ is the cross-ratio function (Oakes, 1989), which measures the strength of association of two correlated failure times. Assume the proband and relative have a common marginal baseline distribution. Then under the class of copula models (Genest and MacKay, 1986; Marshall and Olkin, 1988; Shih and Louis, 1995) in which a bivariate survival distribution is specified in terms of the marginal distributions and a copula function, the conditional hazard function above can be expressed by

$$\lambda(t\mid t_0,\delta_0)=\lambda_0(t)g_\alpha\{\Lambda_0(t),\Lambda_0(t_0)\}\psi_\alpha\{\Lambda_0(t),\Lambda_0(t_0)\}^{\delta_0},\tag{5}$$

where Λ_0 is the baseline cumulative hazard and α is the association parameter characterizing the copula function. Functions g_α and ψ_α depend on t and t_0 through $\Lambda_0(t)$ and $\Lambda_0(t_0)$. The Clayton model considered in the previous section belongs to the class of copula models and is uniquely characterized by the constant cross ratio, i.e., $\psi_\alpha(\Lambda_0(t),\Lambda_0(t_0))=\exp(\alpha)$ for all t,t_0. Its g function is given by

$$
\begin{aligned}
&g_\alpha(\Lambda_0(t),\Lambda_0(t_0))\\
&\quad=\frac{\exp[\Lambda_0(t)\{1-\exp(\alpha)\}]}{\exp[\Lambda_0(t)\{1-\exp(\alpha)\}]+\exp[\Lambda_0(t_0)\{1-\exp(\alpha)\}]-1}.
\end{aligned}
$$

Suppose now the covariates z and z_0 are recorded for the relative and proband, respectively. Assume the proportional hazards model holds for the marginal distribution as in (2). Then the structure in (5) still holds but with $\lambda_0(t)$ replaced by $\lambda_0(t)\exp(\beta' z)$, $\Lambda_0(t)$ by $\Lambda_0(t)\exp(\beta' z)$, and $\Lambda_0(t_0)$ by $\Lambda_0(t_0)\exp(\beta' z_0)$. The conditional hazard in (5) becomes

$$
\begin{aligned}
&\lambda(t\mid t_0,\delta_0,z_0,z)\\
&\quad=\lambda_0(t)\exp(\beta' z)(t)g_\alpha\{\Lambda_0(t)\exp(\beta' z),\Lambda_0(t_0)\exp(\beta' z_0)\}\\
&\qquad\times\psi_\alpha\{\Lambda_0(t)\exp(\beta' z),\Lambda_0(t_0)\exp(\beta' z_0)\}^{\delta_0}.
\end{aligned}
\tag{6}
$$

For the case of the Clayton model, ψ is still equal to $\exp(\alpha)$. Note that the above conditional hazards model allows incorporation of the covariates of both the proband and the relative and the marginal interpretation of the covariate effects is preserved.

3.2 *Estimation*

3.2.1 *Estimating* Λ_0. For $j=1,\ldots,m_i$ and $i=1,\ldots,2n$, let $Y_{ij}(u)=1(t_{ij}\geq u)$, $N_{ij}(u)=\delta_{ij}1(t_{ij}\leq u)$, $g_\alpha^{ij}(\Lambda_0,u)=g_\alpha\{\Lambda_0(u)\exp(\beta' z_{ij}),\Lambda_0(t_{i0})\exp(\beta' z_{i0})\}$, and $\psi_\alpha^{ij}(\Lambda_0,u)=\psi_\alpha\{\Lambda_0(u)\exp(\beta' z_{ij}),\Lambda_0(t_{i0})\exp(\beta' z_{i0})\}$. Also let

$$
\begin{aligned}
&S^{(0)}(\beta,\alpha,\Lambda_0,u)\\
&\quad=\sum_{i=1}^{2n}\sum_{j=1}^{m_i}Y_{ij}(u)\exp(\beta' z_{ij})g_\alpha^{ij}(\Lambda_0,u)\psi_\alpha^{ij}(\Lambda_0,u)^{\delta_{i0}}.
\end{aligned}
$$

The structure of (6) suggests that we can estimate Λ_0 in the spirit of the Nelson–Aalen estimator. We treat $\exp(\beta' z_{ij})g_\alpha^{ij}(\Lambda_0,t_{ij})\psi_\alpha^{ij}(\Lambda_0,t_{ij})^{\delta_{i0}}$ as the (time-dependent) risk score for the jth relative in the ith family, and $S^{(0)}(\beta,\alpha,\Lambda_0,u)$ is the sum of the risk scores of the relatives from all the families at time u. Then for a fixed value of (β,α), an analog to the Nelson–Aalen estimator for Λ_0 is the solution to

$$\Lambda_0(t)=\int_0^t\frac{1}{S^{(0)}(\beta,\alpha,\Lambda_0,u)}\,dN_{++}(u),\tag{7}$$

where $N_{++}(u)=\Sigma_{i=1}^{2n}\Sigma_{j=1}^{m_i}N_{ij}(u)$. Note that, since $S^{(0)}$ involves the unknown baseline cumulative hazard Λ_0, iteration is required to solve (7) for Λ_0.

3.2.2 *Estimating α and β.* We begin by treating β as fixed and consider a quasi-partial-likelihood approach for estimating α, in which only the data on (relative, proband) pairs contribute to the estimation and the higher order correlation is ignored. With the risk score $\exp(\beta' z_{ij}) g_\alpha^{ij}(\Lambda_0, t_{ij}) \times \psi_\alpha^{ij}(\Lambda_0, t_{ij})^{\delta_{i0}}$ for each relative, based on the same argument as Cox (1972), we can form the partial likelihood for α as

$$\hat{L}_p = \prod_{i=1}^{2n} \prod_{j=1}^{m_i} \left\{ \frac{\exp(\beta' z_{ij}) g_\alpha^{ij}(\hat{\Lambda}_0, t_{ij}) \psi_\alpha^{ij}(\hat{\Lambda}_0, t_{ij})^{\delta_{i0}}}{S^{(0)}(\beta, \alpha, \hat{\Lambda}_0, t_{ij})} \right\}^{\delta_{ij}}. \tag{8}$$

The $\hat{}$ over L_p is used to indicate that an estimate of Λ_0 is inserted in the likelihood. The estimate of α is obtained by maximizing $\hat{L}_p$ with respect to α.

An alternative approach for estimating α that we will evaluate is to incorporate all the correlation information in the likelihood. Specifically, we can estimate α using the full likelihood but with the marginal baseline distribution estimated by (7). For example, under the Clayton model, following (4), we can estimate the scalar α by maximizing the following quasi-likelihood:

$$\begin{aligned} \hat{L} \propto & \prod_{i=1}^{2n} \left[\sum_{j=0}^{m_i} \exp\{-\hat{\Lambda}_0(t_{ij})(1-\theta)\exp(\beta' z_{ij})\} - m_i \right]^{\frac{1}{(1-\theta)} - d_i} \\ & \times \left[\prod_{j=1}^{d_i} \{\theta(j-1) + 2 - j\} \right] \\ & \times \left[\prod_{j=0}^{m_i} [\exp\{-\hat{\Lambda}_0(t_{ij})(1-\theta)\exp(\beta' z_{ij})\} \exp(\beta' z_{ij})]^{\delta_{ij}} \right]. \end{aligned} \tag{9}$$

By incorporating the whole correlation structure, $\hat{L}$ should be more efficient in estimating the association parameter α than $\hat{L}_p$. Computationally, however, $\hat{L}_p$ is simpler. Besides, if the familial association varies over pairs, then L_p is readily extended to incorporate the pair-dependent association, but constructing a joint distribution for $\hat{L}$ would be complex.

Up till now, β has been treated as fixed. Suppose now we are interested in estimating both β and α simultaneously. Either $\hat{L}_p$ or $\hat{L}$ can be maximized to obtain an estimate of β. However, both likelihoods ignore the contribution from the probands data. Similar to Li et al. (1998), we will include the conditional likelihood L_1^c in the estimation of β. Thus, we estimate β by maximizing either $\hat{L}_p L_1^c$ or $\hat{L} L_1^c$.

Because estimating Λ_0 and (β, α) requires knowledge of the other component, iteration is needed in finding the solution $(\hat{\Lambda}_0, \hat{\alpha}, \hat{\beta})$. With an initial guess $(\hat{\Lambda}_0^{(0)}, \hat{\alpha}^{(0)}, \hat{\beta}^{(0)})$, we update $\hat{\Lambda}_0$ by

$$\hat{\Lambda}_0^{(\nu+1)}(t) = \int_0^t \frac{1}{S^{(0)}(\hat{\beta}^{(\nu)}, \hat{\alpha}^{(\nu)}, \hat{\Lambda}_0^{(\nu)}, u)} dN_{++}(u). \tag{10}$$

We keep updating $\hat{\Lambda}_0$ using (10) until convergence. Then with the current estimate of Λ_0, we maximize either $\hat{L}_p L_c$ or $\hat{L} L_c$ with respect to (β, α) to update $(\hat{\beta}^{(\nu)}, \hat{\alpha}^{(\nu)})$. With the updated estimate for (β, α), we go back to (10), and iteration proceeds until $(\hat{\Lambda}_0, \hat{\beta}, \hat{\alpha})$ reaches convergence. Further work is needed to develop the asymptotic theory for the semiparametric estimation method we described. In our illustration shown later, we use the bootstrap method to obtain the standard errors.

4. Simulation Study

We conducted a simulation study to evaluate the performance of the proposed method. In the following, we describe how to generate a sample of case–control family data. We first generated the correlated failure times for each family from the Clayton model, in which the marginal distribution follows a Weibull model with baseline survival function $S_0(t) = \exp\{-(.013t)^{5.148}\}$ and the cross ratio $\exp(\alpha)$ ranges from 1.5 to 3 (α ranges from .406 to 1.10). The censoring time of each individual is independent of the age of onset and follows a normal distribution with mean 65 and standard deviation 12, which results in approximately 40% censoring. Ages of onset/censoring times were generated for 10,000 families. Then 200 case probands were randomly selected from the pool of 10,000 families. Each case proband is matched with a control proband within 1-year of age. Once these probands are identified, data on their relatives are included. There are a total of 400 families. The simulation was repeated 500 times for each scenario considered.

We fit the data using the proposed method assuming the Clayton model. We compare the proposed method with three alternative semiparametric methods for our problem considered by Hsu et al. (1999): the Cox model, a stratified Cox model, and the pseudo-partial-likelihood approach of Hsu et al. (1999).

The first method (Cox model) is a proportional hazards model for the age of onset of the relatives with their proband's age and disease status treated as the covariates, i.e.,

$$\lambda_{ij}(t \mid t_{i0}, \delta_{i0}) = \lambda_0(t) \exp(\beta_0 t_{i0} + \alpha \delta_{i0} + \beta' z_{ij}). \tag{11}$$

The second method (stratified Cox model) is based on the stratified conditional hazard given by

$$\lambda_{ij}(t \mid t_{i0}, \delta_{i0}) = \lambda_{0i}(t) \exp(\alpha \delta_{i0} + \beta' z_{ij}). \tag{12}$$

We fit this model by stratifying on the proband's integer age. All the individuals with the same integer age of proband form the same stratum.

The third method (pseudo-partial-likelihood) assumes the stratified Cox model (12) holds and forms a pseudo-partial-likelihood for β and α by comparing the risk scores $R_{ij} = \exp\{\alpha \delta_{i0} + \beta' z_{ij}\}$ among all the pairs of relatives of different families in each matched set. The likelihood has the following representation,

$$\begin{aligned} & L_{\text{ppl}}(\beta, \alpha) \\ & = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \prod_{k=1}^{m_{n+i}} \left\{ \frac{R_{ij}}{R_{ij} + I(t_{i+n,k} \geq t_{ij}) R_{i+n,k}} \right\}^{\delta_{ij}} \\ & \qquad \times \left\{ \frac{R_{i+n,k}}{R_{i+n,k} + I(t_{ij} \geq t_{i+n,k}) R_{ij}} \right\}^{\delta_{i+n,k}}. \end{aligned}$$

Note that the marginal hazard (2) is a special case of the conditional hazards (11) and (12) by letting $t_{i0} = 0$ and $\delta_{i0} = 0$. However, the interpretation of β in (11) and (12) is conditional on the proband's survival data. Besides, it is

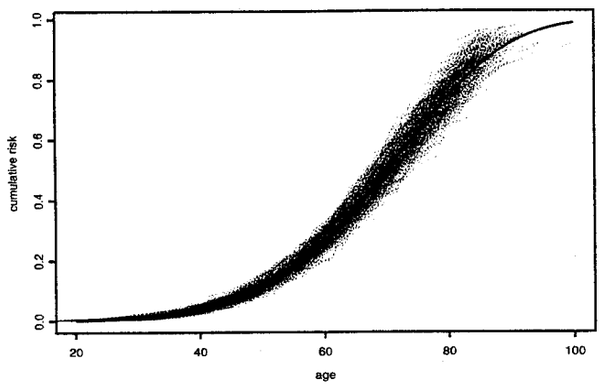

Figure 1. Cumulative risk vs. age of onset. Solid line: true distribution; dotted lines: estimated distributions from 250 simulations.

not clear how one incorporates the covariates of the proband in these approaches. Furthermore, if the disease status of the proband, δ_0, has a proportional effect on the conditional hazard of the relative as specified, then the only bivariate model for continuous data satisfying such a property is the Clayton model (Clayton, 1978), in which the proband's age cannot be factored out from the relative's age as specified in (11).

We first consider bivariate data where each proband has only one relative. Figure 1 exhibits the estimated baseline survival function estimated from the proposed approach with $\exp(\alpha) = 2$, where the dotted lines are estimated distributions from 250 simulations and the solid line is the true distribution. It shows that the marginal baseline distribution is estimated well and there is no apparent bias. Table 1 summarizes the simulation result for the parameter estimators. The upper panel presents the simulation result when there are no covariates. The relative bias is defined as the difference between

Table 1
Simulation results for bivariate data generated from the Clayton model

	Proposed method	Cox	Stratified Cox	Pseudo-partial-likelihood
Without Covariates				
α (= .406)				
Relative bias (%)	−.5	−2	−2.7	−1.0
Variance	.016	.027	.040	.048
Efficiency (%)	100	59	40	33
β_0				
Mean		−.011		
Variance		.00004		
α (= .693)				
Relative bias (%)	.9	.2	.7	.0
Variance	.024	.032	.051	.063
Efficiency (%)	100	75	46	37
β_0				
Mean		−.017		
Variance		.00005		
α (= 1.10)				
Relative bias (%)	.5	−.4	.5	−.3
Variance	.021	.029	.044	.063
Efficiency (%)	100	72	47	33
β_0				
Mean		−.031		
Variance		.00003		
With Covariates				
α (= .693)				
Relative bias (%)	.4	−10	−9	−7
Variance	.015	.023	.038	.056
Efficiency (%)	100	64	39	26
β (= 1)				
Relative bias (%)	.4	8	8	10
Variance	.013	.021	.035	.125
Efficiency (%)	100	61	36	10
β_0				
Mean		−.02		
Variance		.00004		

the simulated mean and true parameter value divided by the true value. Efficiency is defined by the ratio of the simulated variance of the proposed method to that of the specific approach. The relative biases for α from all these approaches are close to zero, but the estimators vary in efficiency. As expected, due to high stratification, estimators from stratified Cox model and pseudo-partial likelihood are not efficient in all the α values considered. The Cox model, on the other hand, has about 25–40% efficiency loss and lends itself to difficulty of interpretation, i.e., the resulting conditional hazard, in which the age of onset of the relative does not depend on the age ($\beta_0 \approx 0$) but only on the disease status of the proband ($\alpha > 0$), does not seem plausible under any continuous bivariate distribution with exchangeable margins.

The lower panel presents the simulation result when there is one binary covariate that operates proportionally on the marginal hazard with $\beta = 1$, $\Pr(z = 1) = .3$, and $\alpha = .693$. The estimators of α and β show biases in all except the proposed approach. In addition to biases, there is great efficiency loss in estimating β. For example, the pseudo-partial likelihood approach has a variance about 10 times as large as that of the proposed method.

To compare the performance of the two approaches using the likelihoods (7) and (8), we generated trivariate failure time data from the Clayton model, where each proband has two relatives. As before, we generated data for 200 case families and 200 control families. Table 2 shows the simulation result. Both approaches produce little bias. The approach incorporating all the correlation information (likelihood (8)) is about 20% more efficient than the other one incorporating only the paired correlation.

In addition to comparing the proposed approach with the existing semiparametric approaches, we study the bias likely incurred by using an *ad hoc* parametric model to fit the baseline hazard function. In this simulation, the marginal distribution follows a nonmonotone piecewise exponential model. The values of the hazards are (0.00023, 0.01101, 0.02957, 0.03721, 0.00805, 0.01423, 9.76e−12, 5.57e−11) with cut-off age at (30, 40, 50, 60, 70, 80, 90). These values of the hazards were taken from the hazard estimates from the carriers of the BRCA 1/2 mutations of the Washington Ashkenazi Study (WAS) (Struewing et al., 1997). The association parameter (α) for the Clayton model is equal to .693. The censoring mechanism is kept the same as before. When we assumed the commonly used Weibull model for the baseline hazard, the estimate of the association parameter failed to converge. With the exponential model, the parameter estimate failed to converge in 20% of the replications. For the rest of the 80% of the replications, the mean of the estimate of α is equal to 1.655. Finally, with the piecewise exponential model with cut-off age at (30, 50, 70, 90), the estimate of the association parameter converged in most (96%) of the replications, with the mean of estimate equal to .756. As expected, when the assumed model is very different from the true model, as in the case of the Weibull and exponential models, the estimator either fails to converge or, if it converges, has a large bias. When the assumed model is not so different from the true model, the bias is smaller, as seen in the final model. The final model could be further extended to mimic the true model. However, the estimation would become more difficult and likely unstable because it involves simultaneous estimation of many parameters in the marginal distribution and the association parameter. In this case, computation of our proposed semiparametric approach would be simpler.

Table 2
Simulation results for trivariate data generated from the Clayton model ($\alpha = .693$)

	Approach based on (7)	Approach based on (8)
Relative bias (%)	.3	−.1
Variance	.015	.012
Efficiency (%)	100	122

The simulation study demonstrates that the proposed method produces little bias and is highly efficient. In addition, unlike the other semiparametric approaches, the proposed method can naturally incorporate covariates for both the relatives and proband in the marginal hazard.

5. Illustration

As an illustration, we construct a case–control family data set from the WAS study. In this study, more than 5000 volunteer Ashkenazi Jews living in the Washington, D.C., area provided blood samples for genotyping of BRCA1/BRCA2 mutations. They also gave family history information on breast and other common cancers. We use a subset of the data that contain (mother, daughter) pairs where daughters are noncarrier volunteers (without any BRCA1/BRCA2 mutations). There are 193 cases of breast cancer in the volunteers. We matched these case volunteers with control volunteers on the onset ages within 5 years. Hence, the data set used for illustration contains 386 mother–daughter pairs. Among the mothers, 15.3% had breast cancer at some time. Covariates included in the analysis are age of first birth (AFB) and parity ('1' = 1 or 2 children, '0' $\geq$ 3 children).

Although in this study only the volunteers were genotyped and no DNA samples were available on their relatives, Struewing et al. (1997) originally estimated disease risk from the mutations from the disease history of the relatives of the participating volunteers and not from the volunteers themselves. It was argued that the volunteers may have a strong survival effect on their participation in the study as a diseased individual could participate in the study only if she remained alive until the study took place. Mothers of the volunteers, however, were immune to this kind of bias because their data were collected through the volunteers and a diseased mother could be included in the study even if she had died from the disease before the study took place. For the same reason, we did not include the conditional likelihood L_1^c of the case–control sample of the probands in our analysis and estimated the covariate effects and familial association solely from the quasi-partial-likelihood, $\hat{L}_p$, which conditions on all the data available from the volunteers. Standard errors of the estimates were obtained using the bootstrap method, where 500 bootstrap samples were drawn from mother–daughter pairs with replacement.

The estimation result is presented in Table 3. Although the size of familial association seems modest (cross ratio = $\exp(\alpha)$

Table 3
Estimation results for a case–control sample from the WAS study

$\lambda(t) = \lambda_0(t)\exp(\beta_1 \text{parity} + \beta_2 \text{AFB}/10)$		
	Estimate	Standard error
β_1	.11	.28
β_2	.19	.26
α	.37	.26

= 1.45) compared with the well-known effect of a strong familial component of the disease, it is consistent with the report by Kaufman and Struewing (1999), who estimated familial correlation as the effect of family history on the risk of the disease in the noncarrier volunteers as measured by odds ratio (estimated as 1.5) in a logistic regression model. Closeness of the estimates from the two approaches is not surprising given that, in both approaches, information on familial correlation comes from the link between volunteer and their relatives. Chatterjee et al. (2001), on the other hand, estimated familial correlation using links between the relatives of the same volunteer and found a stronger familial correlation (cross ratio ≈ 2). Although not addressed in this article, the apparent inconsistency in the magnitude of familial correlation as estimated by two different types of data from the same study poses interesting epidemiologic questions.

The point estimates corresponding to the regression effect of both age at first birth and parity are consistent with other studies (Chie et al., 2000). Older age at first birth and fewer children both are associated with increasing risk of breast cancer in women. The relative magnitude of the regression coefficients suggest that, among women who have at least one child, age at first birth is more important than the number of children for reducing the risk of the disease. Finally, we note that, although the parameter estimates obtained from our analysis seem to be consistent with other studies, none of the estimates are statistically significant due to the small size of our data, which has a low cancer rate for the mothers.

6. Discussion

We have proposed a quasi-partial-likelihood approach for estimating the effects of covariates on the age of onset and familial association in case–control family studies. Our methods take into account the retrospective nature of the case–control study design and use an association structure between ages of onset for all members of the same family to obtain the estimate of the marginal baseline distribution of age of onset. Our simulation study shows that, compared with the existing semiparametric approaches, the proposed estimators have little bias and are highly efficient.

Although the proposed method is postulated under a one-to-one match setting, it is readily generalized to other case–control study settings, such as multiple control families matched to one case family. Development of the asymptotic theory for the proposed estimators is a topic of ongoing research. It involves modern theory on adaptive estimation for semiparametric models with correlated data (Bickel et al., 1993; Murphy, 1994, 1995; van der Vaart and Wellner, 1996). Finally, the proposed method requires that the covariate information is available on both the case–control sample and their relatives. In many applications, however, it may be difficult or cost prohibitive to obtain covariate information on some or any relatives of the cases and controls. In this situation, the covariate data on relatives can be treated as missing. Future research is merited on extending the proposed method to account for the possibility of missing covariates.

Acknowledgement

The authors would like to thank Dean Follmann and Mitch Gail for their helpful comments.

Résumé

Dans les études cas-témoins familiales dont le critère de jugement est la survie, on peut utiliser l'âge au début de la maladie pour établir l'existence d'une aggravation familiale de maladie et la relation entre la maladie et les facteurs de risques génétiques ou environnementaux. Du fait du caractère rétrospectif des enquêtes cas-témoins, les méthodes établies pour l'analyse des données de survie corrélées recueillies prospectivement ne s'appliquent pas directement. Dans cet article, nous proposons une approche semi paramétrique de quasi vraisemblance partielle, pour estimer simultanément l'effet des covariables sur l'âge au début de la maladie et l'association des âges au début parmi les membres de famille, qui ne nécessite pas de spécifier la distribution marginale de base. Nous avons réalisé une étude de simulation pour évaluer les performances de la méthode proposée et la comparer aux méthodes semi-paramétriques qui existent. Ces résultats de simulation montrent que cette nouvelle approche a des performances meilleures en terme d'efficacité et de conséquence. Nous illustrons la méthode à partir d'un sous ensemble de données extraites de l'étude Washington Ashkenazi.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, Maryland: Johns Hopkins University Press.

Chatterjee, N. and Shih, J. (2001). A bivariate cure-mixture approach for modeling familial association in diseases. *Biometrics* **57,** 779–786.

Chatterjee, N., Shih, J., Hartge, P., Brody, L., Tucker, M., and Wacholder, S. (2001). Association and aggregation analysis using kin-cohort designs with applications to genotype and family history data from the Washington Ashkenazi Study. *Genetic Epidemiology* **21,** 123–138.

Chie, W. C., Hsieh, C., Newcomb, P. A., Longnecker, M. P., Mittendorf, R., Greenberg, E. R., Clapp, R. W., Burke, K. P., Titus-Ernstoff, L., Trentham-Dietz, A., and MacMahon, B. (2000). Age at any full-term pregnancy and breast cancer risk. *American Journal of Epidemiology* **151,** 715–722.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65,** 141–151.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Genest, C. and MacKay, R. J. (1986). The joy of copulas: Bivariate distributions with given marginals. *The American Statistician* **40,** 280–283.

Hsu, L., Prentice, R. L., Zhao, L. P., and Fan, J. J. (1999). On dependence estimation using correlated failure time data from case–control family studies. *Biometrika* **86,** 743–753.

Kaufman, D. J. and Struewing, J. (1999). Re: Effect of BRCA1 and BRCA2 on the association between breast cancer risk and family history. *Journal of the National Cancer Institute* **91,** 1250.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of age of onset data from case–control family studies. *Biometrics* **54,** 1030–1339.

MacLean, C. J., Neale, M. C., Meyer, J. M., and Kendler, K. S. (1990). Estimating familial effects on age of onset and liability to schizophrenia. II. Adjustment for censored data. *Genetic Epidemiology* **7,** 419–426.

Marshall, A. W. and Olkin, I. (1988). Families of multivariate distributions. *Journal of the American Statistical Association* **83,** 834–841.

Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics* **22,** 712–731.

Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23,** 182–198.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84,** 487–493.

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time model. *Biometrika* **65,** 153–158.

Prentice, R. L. and Hsu, L. (1997). Estimating equations for hazard ratio and correlation parameters in multivariate failure time analysis. *Biometrika* **84,** 131–145.

Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51,** 1384–1399.

Struewing, J. P., Hartge, P., Wacholder, S., Baker, S. M., Berlin, M., McAdams, M., Timmerman, M. M., Lawrence, B. C., and Tucker, M. A. (1997). The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *The New England Journal of Medicine* **336,** 1401–1408.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* New York: Springer-Verlag.

Whittemore, A. S. (1995). Logistic regression of family data from case–control studies. *Biometrika* **82,** 57–67.

Received March 2001. *Revised April* 2002.
Accepted May 2002.

Appendix

Derivation of (4). The joint survival function of the proband and m_i members of family i is given by

$$S(t_{i0}, \ldots, t_{im_i} \mid \boldsymbol{Z}_i) = \left[\sum_{j=0}^{m_i} \exp\{-\Lambda_0(t_{ij};\gamma)(1-\theta)\exp(\beta' z_{ij})\} - m_i \right]^{\frac{1}{(1-\theta)}}.$$

Taking the derivative of $S(t_{i0}, \ldots, t_{im_i} \mid \boldsymbol{Z}_i)$ with respect to t_{ij}'s such that $\delta_{ij} = 1$ yields the product of the first three components in (4).

The last component,

$$\exp\{\Lambda_0(t_{i0};\gamma)\exp(\beta' z_{i0})\}\{\exp(\beta' z_{i0})\lambda_0(t_{i0};\gamma)\}^{-\delta_{i0}},$$

is the reciprocal of the likelihood contribution of the proband, which is needed because L_2 is the conditional likelihood of the family survival data given the proband survival data.